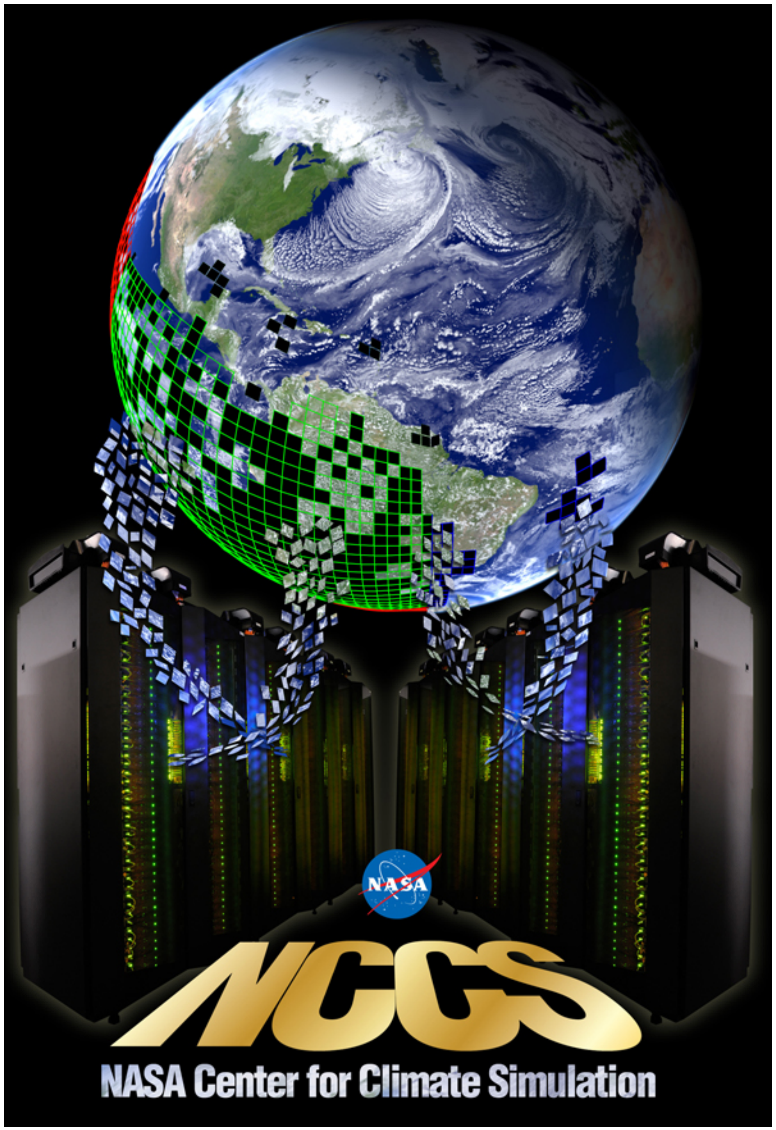


User Forum

NASA Center for Climate Simulation

High Performance Science

October 18, 2016

Agenda

- Staff Additions
- User Survey Responses
- Utilization and Availability
- Architecture Evolution
- FY17 Hardware Updates and Procurements
- NCCS Operational Updates
- NCCS User Recognitions

Staff Additions

Welcome to New Members of the NCCS and CISTO Team:

Jordan Caraballo-Vega/NCCS Intern, Security

Aaron Knister/CSRA, System Administration

Jonathan Mills/CSRA, System Administration

Jian Li/CSRA, Data Services

Mahdi Magrhebi/CSRA, Technical User Services

Carrie Spear/CSRA, HPC Architect

Jim Shute/CSRA, Data Services (ArcGIS Expert)

Dan'l Pierce/CSRA, Deputy Program Manager

Bob Peirce/NASA, CISTO Associate Chief

Nancy Carney/GST, HEC Allocation Specialist

NASA High End Computing (HEC) November Allocation Reminders

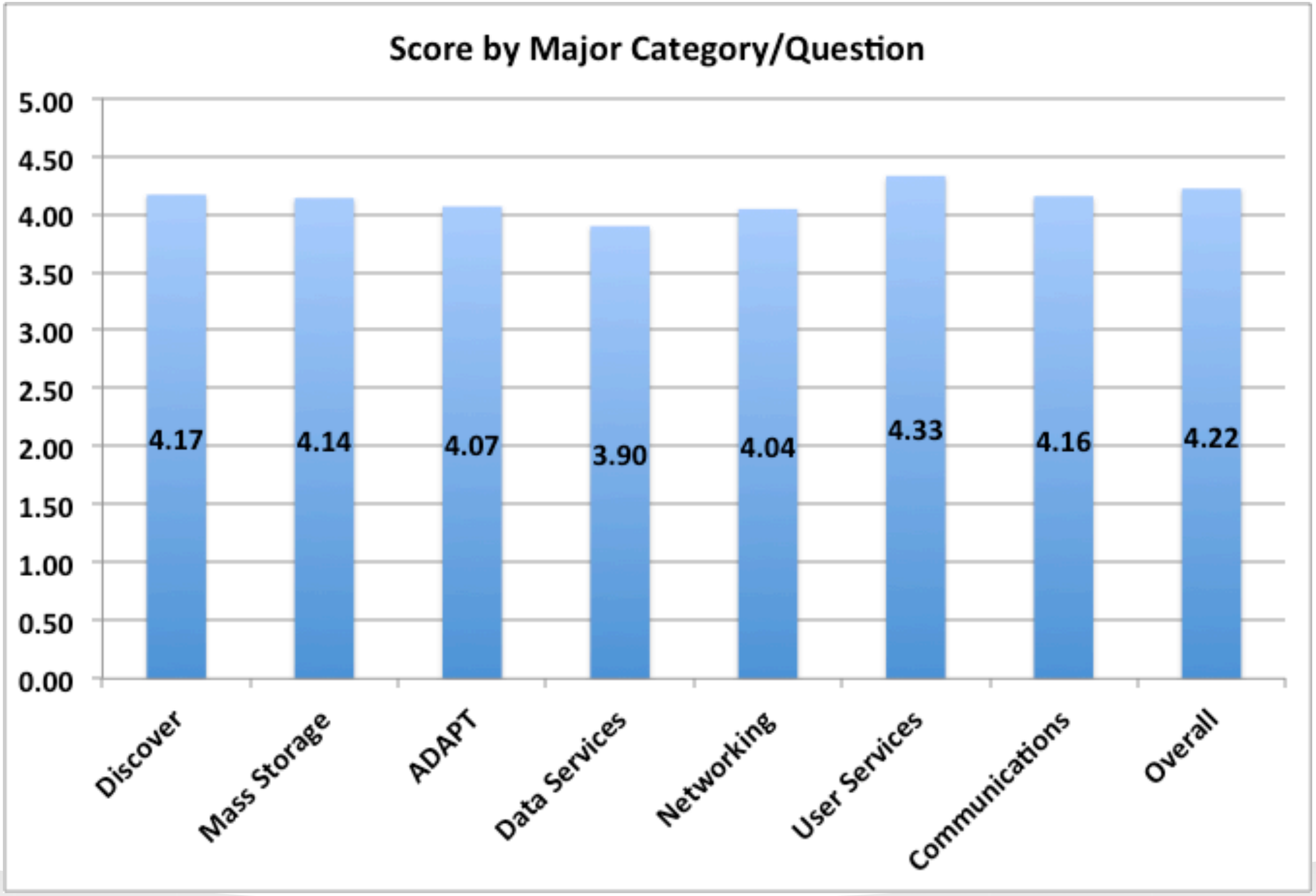
- Deadline for submitting requests was September 20th, 2016
- ALL allocations expire October 31, 2016
  - Both November 2015 and May 2016 allocations expire
- New allocation period is for November 1, 2016 through September 30, 2017
  - Note that this is an 11 month period
- Allocation requests received so far have been submitted to HQ for review
  - Expect these to be finalized next week
- If you have not submitted an eBooks request, talk to Nancy Carney (nancy.s.carney@nasa.gov).
  - There still may be time to get an allocation, but you must act quickly.

User Survey Responses

Dan Duffy,
HPC Lead and NCCS Lead Architect

2016 NCCS User Survey Scoring

- Scoring
  - 5 – Excellent
  - 4 – Very Good
  - 3 – Good
  - 2 – Fair
  - 1 – Poor



Open Ended Responses about Outstanding Service

- Everything involved with my user experience has been outstanding
- Responsiveness when problems do arise
- The people @ NCCS are outstanding.
- Support. Fast and reliable, great job.
- The professional dedicated support we receive has got to top the list.
- The staff and their dedication to resolve problems.

The user community consistently wrote about the outstanding support and personnel within the NCCS!

Open Ended Responses about Future Needs

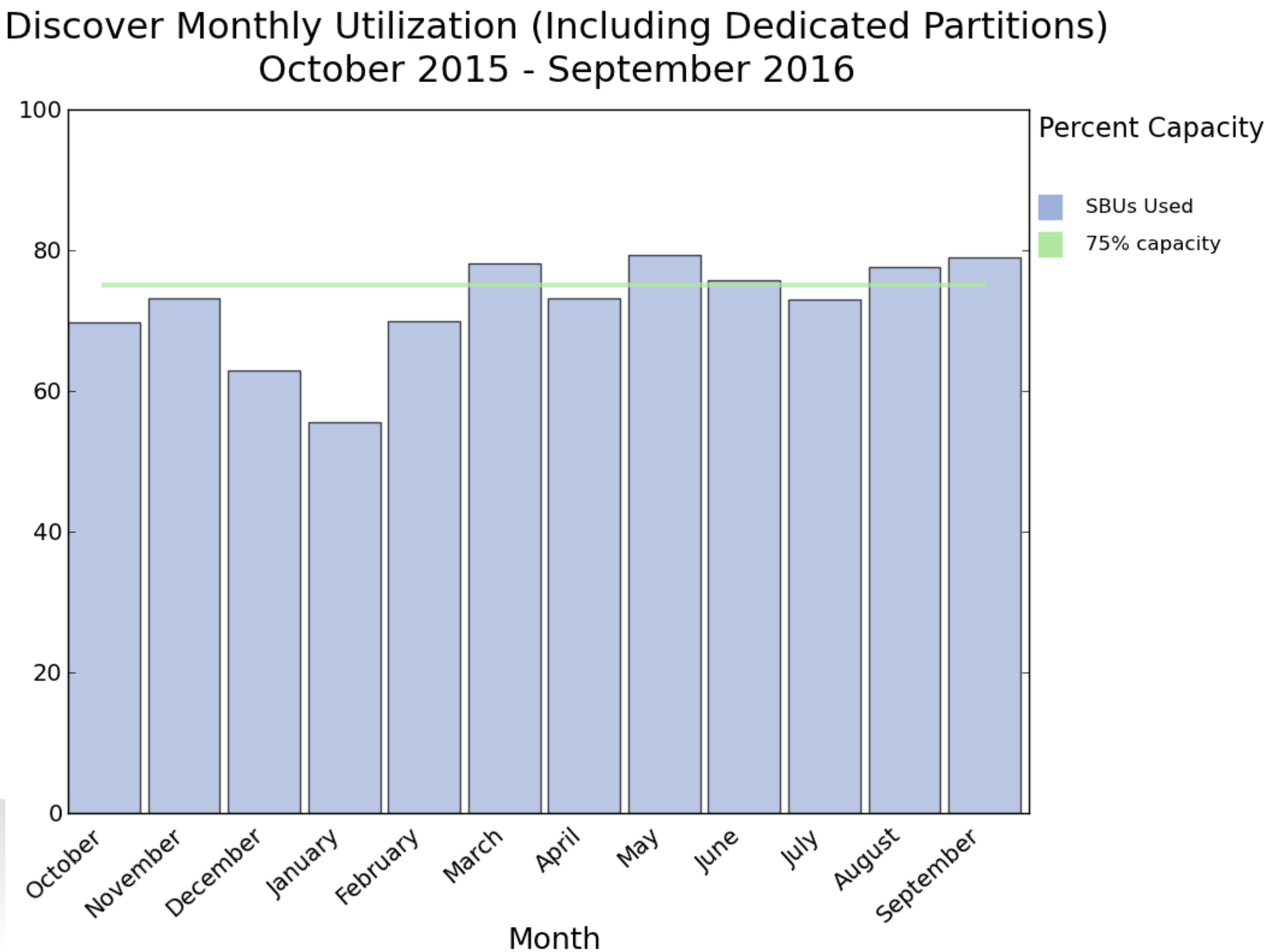
- Not just me, but the whole GISS community, will be needing many more CPUs.
- More complex data assimilation will require more processing power and more storage Increasing need for CPU and disk space.
- More dedicated user storage space.
- Will need faster networks and larger disk/memory allocations to handle exponentially larger data sets.
- We will need more resources as our models go to higher resolution.
- While we are currently running large jobs infrequently, I think this will switch to smaller jobs much more frequently over the next few years.
- NCCS provides essential service for the accomplishment of my scientific goals. They are absolutely essential, and this is why I (and I suppose many others) react so strongly when service is interrupted.
- More data storage required. I expect to add more routine runs using LIS and NUWRF.
- More of everything.

The general theme for future needs by our current user community is more of everything: processors, storage, networking, etc.

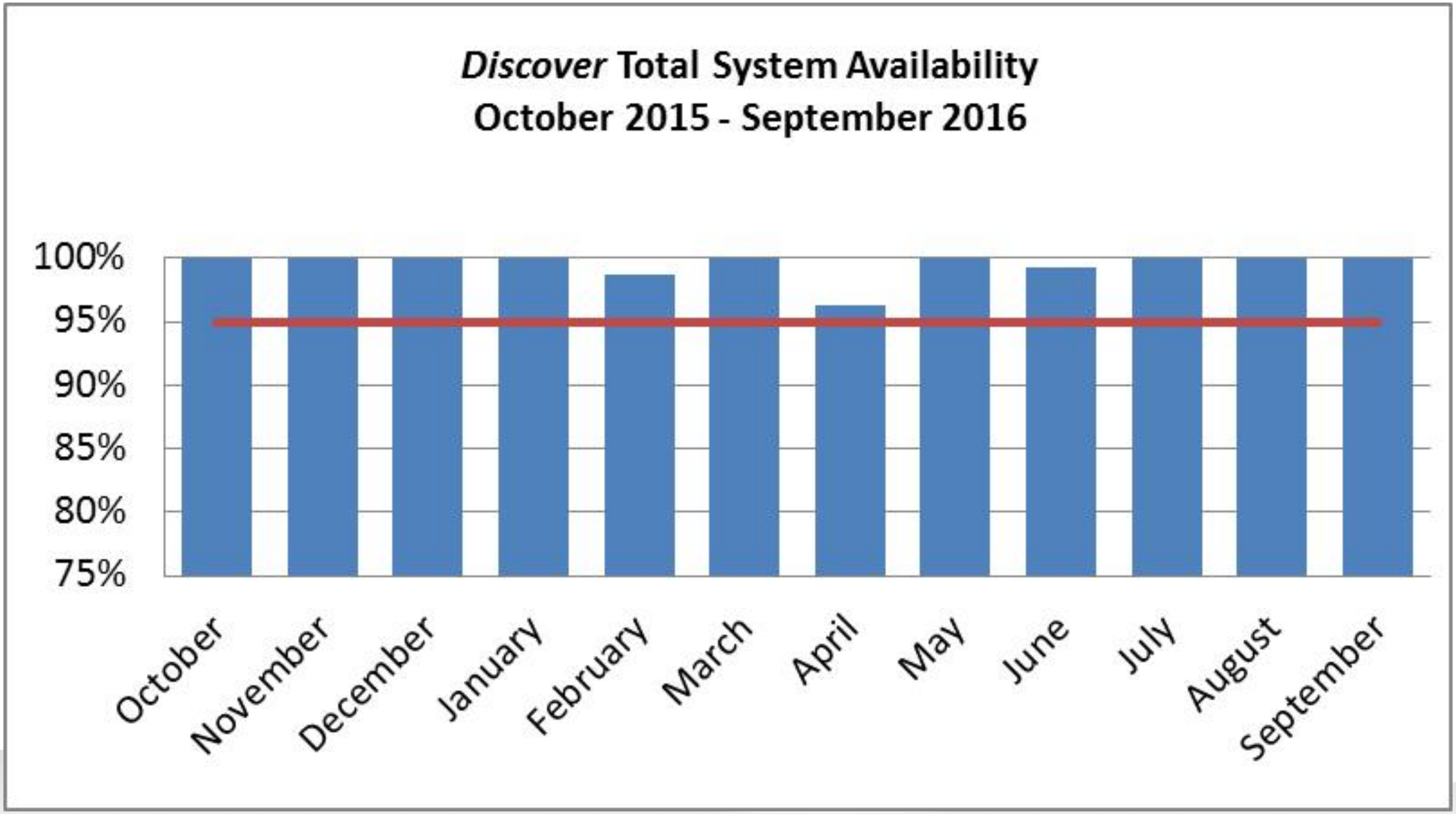
Utilization and Availability

Dan Duffy,
HPC Lead and NCCS Lead Architect

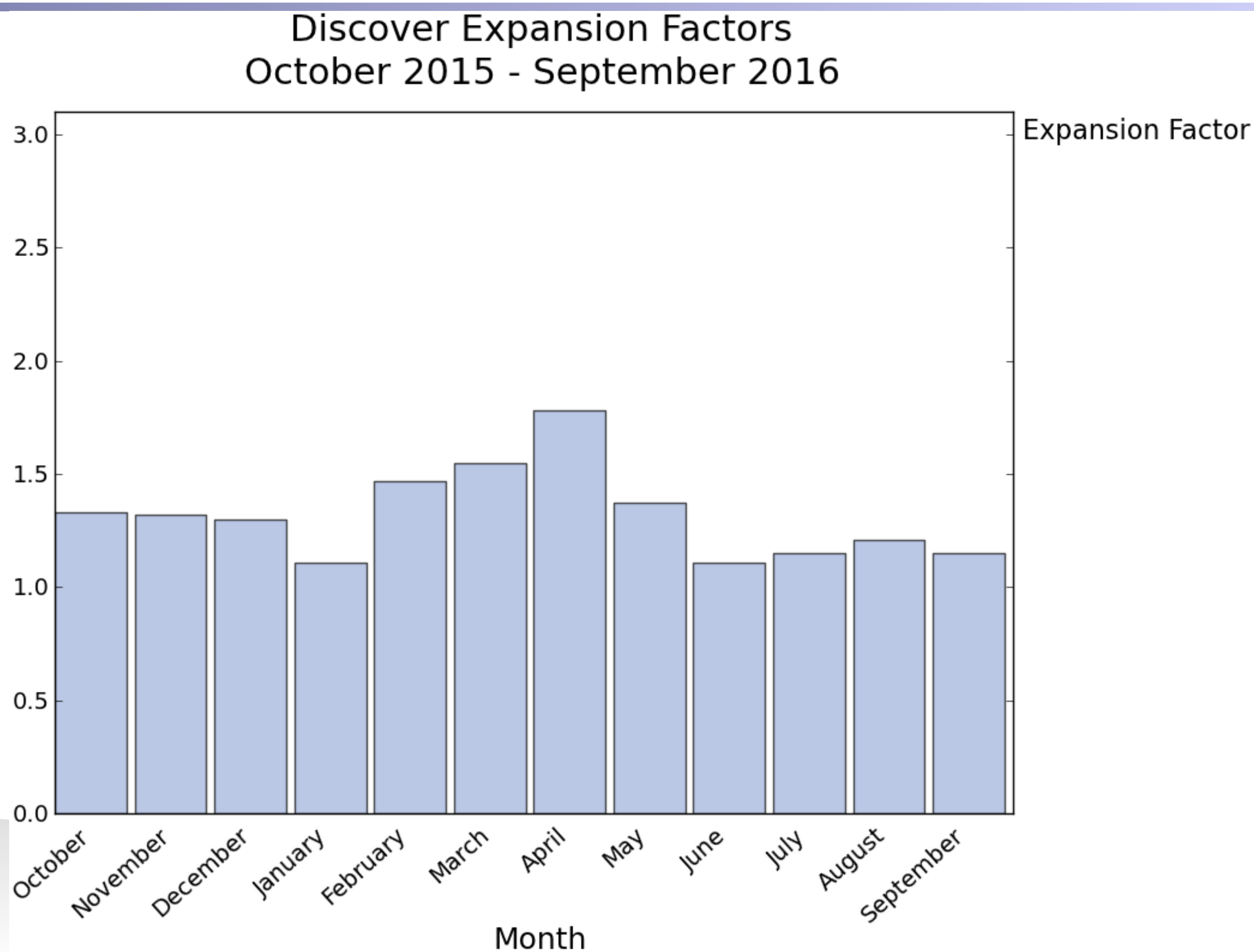
Discover 12-Month Utilization Percentage Trend



Discover System Availability



Discover Expansion Factors – 12-Month Trend



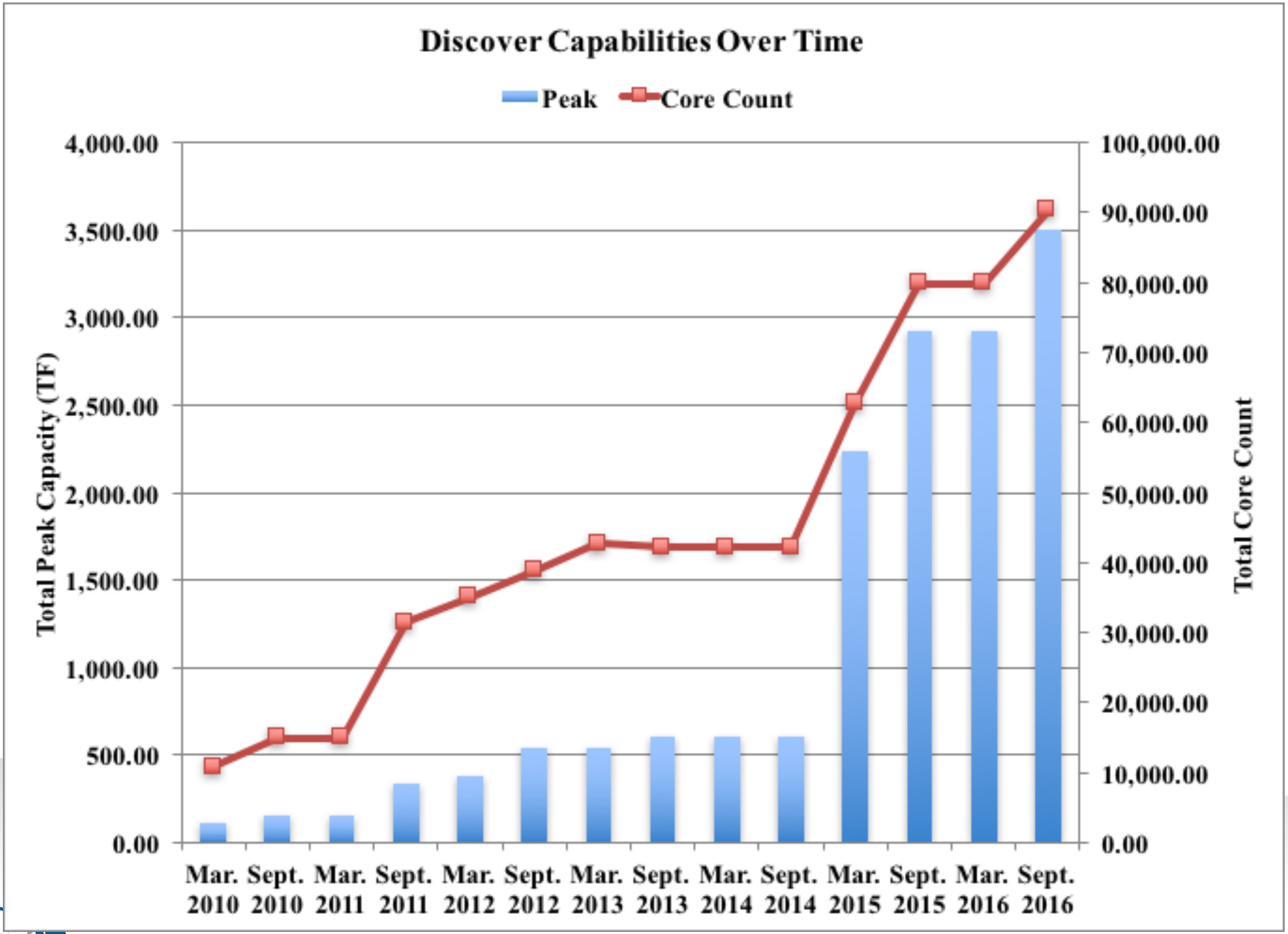
NCCS Architecture Evolution

Dan Duffy,
HPC Lead and NCCS Lead Architect

Discover Scalable Unit Evolution

2014	SCU1 258 Nodes 3,096 Cores 2.8 GHz 24 GB Westmere DDR IB 34.7 TF	SCU2 258 Nodes 3,096 Cores 2.8 GHz 24 GB Westmere DDR IB 34.7 TF	SCU3 258 Nodes 3,096 Cores 2.8 GHz 24 GB Westmere DDR IB 34.7 TF	SCU4 258 Nodes 3,096 Cores 2.8 GHz 24 GB Westmere DDR IB 34.7 TF	SCU7 1,200 Nodes 14,400 Cores 2.8 GHz 24 GB Westmere QDR IB 161.3 TF	SCU8 480 Nodes 7,680 Cores 2.6 GHz 32 GB Westmere QDR IB Xeon Phi 160.0 TF	SCU9 480 Nodes 7,680 Cores 2.6 GHz 64 GB Westmere FDR IB Xeon Phi 160.0 TF
2015	SCU11 612 Nodes 17,136 Cores 2.6 GHz 128 GB Haswell FDR IB 712,858 TF	SCU12 612 Nodes 17,136 Cores 2.6 GHz 128 GB Haswell FDR IB 712,858 TF		SCU10 1,080 Nodes 30,240 Cores 2.6 GHz 128 GB Haswell FDR IB 1,257,984 TF		SCU8 480 Nodes 7,680 Cores 2.6 GHz 32 GB Westmere QDR IB Xeon Phi 160.0 TF	SCU9 480 Nodes 7,680 Cores 2.6 GHz 64 GB Westmere FDR IB Xeon Phi 160.0 TF
2016	SCU11 612 Nodes 17,136 Cores 2.6 GHz 128 GB Haswell FDR IB 712,858 TF	SCU12 612 Nodes 17,136 Cores 2.6 GHz 128 GB Haswell FDR IB 712,858 TF		SCU10 1,080 Nodes 30,240 Cores 2.6 GHz 128 GB Haswell FDR IB 1,257,984 TF	SCU13 648 Nodes 18,144 Cores 2.6 GHz 128 GB Haswell FDR IB 754,790 TF		SCU9 480 Nodes 7,680 Cores 2.6 GHz 64 GB Westmere FDR IB Xeon Phi 160.0 TF

Computing Capacity Evolution

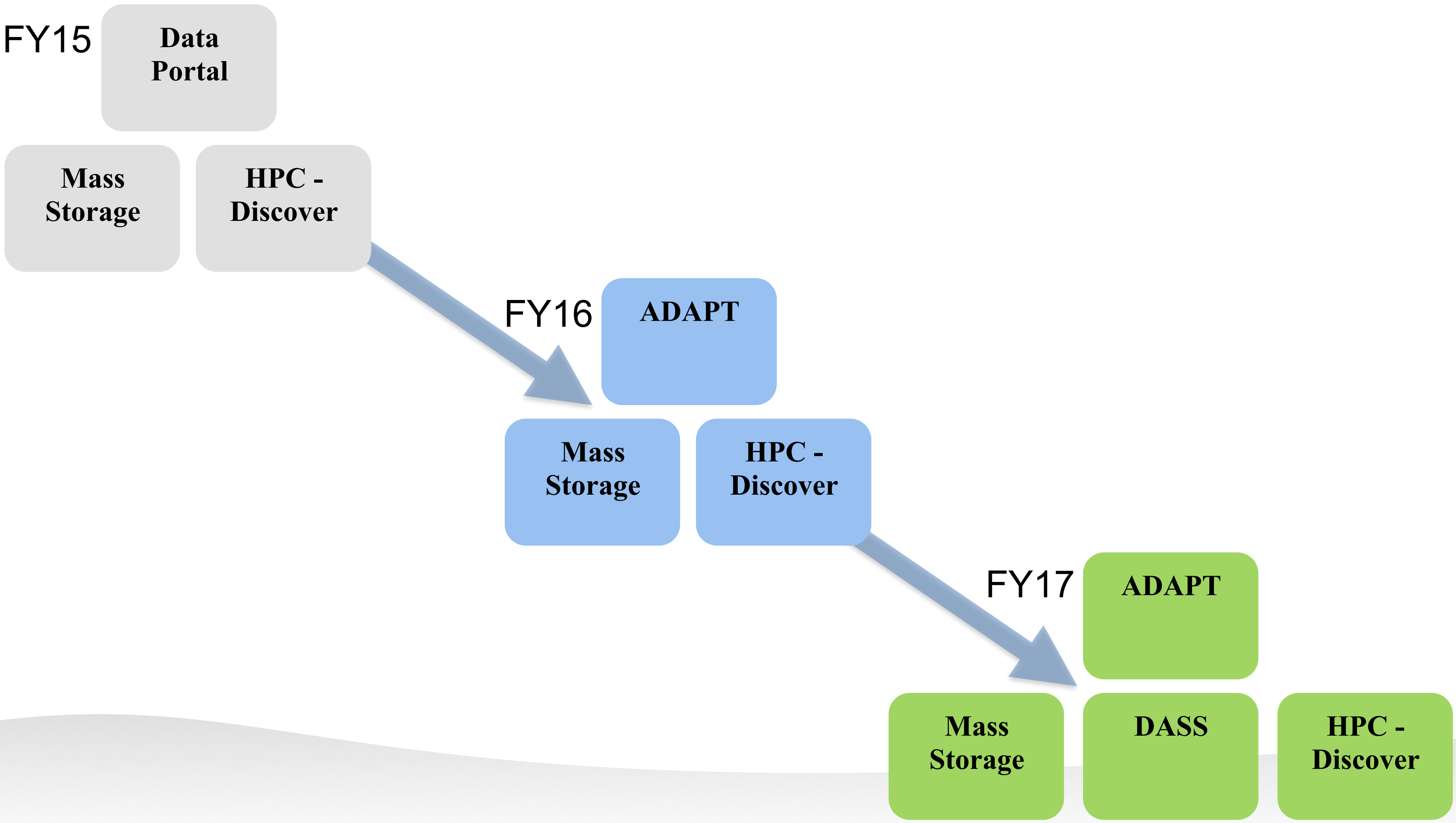


Discover Scratch Disk Evolution

Calendar	Description	Decommission	Total Usable Capacity (TB)
2012	Combination of DDN disks	None	3,960
Fall 2012	NetApp1: 1,800 by 3 TB Disk Drives; 5,400 TB RAW (prior to RAID protection)	None	9,360
Fall 2013	NetApp2: 1,800 by 4 TB Disk Drives; 7,200 TB RAW (prior to RAID protection)	None	16,560
Early 2015	DDN10: 1,680 by 6 TB Disk Drives, 10,080 TB RAW (prior to RAID protection)	DDNs 3, 4, 5	~26,000
Mid 2015	DDN11: 1,680 by 6 TB Disk Drives, 10,080 TB RAW (prior to RAID protection)	DDNs 7, 8, 9	~33,000
Mid 2016	**DDN12: 1,680 by 6 TB Disk Drives, 10,080 TB RAW (prior to RAID protection)**	**None**	**~40,000**
Early 2017	**13+ PB RAW (prior to RAID protection)**	**TBD**	**~50,000**

- Usable capacity differs from raw capacity for two reasons. First, the NCCS uses RAID6 (double parity) to protect against drive failures. This incurs a 20% overhead for the disk capacity. Second, the file system formatting is estimated to also need about 5% of the overall disk capacity. The total reduction from the RAW capacity to usable space is about 25%.

Evolution of Systems



Advanced Data Analytics Platform (ADAPT)

Dan Duffy,
HPC Lead and NCCS Lead Architect

Advanced Data Analytics Platform (ADAPT) Platform-as-a-Service (PaaS) Architecture

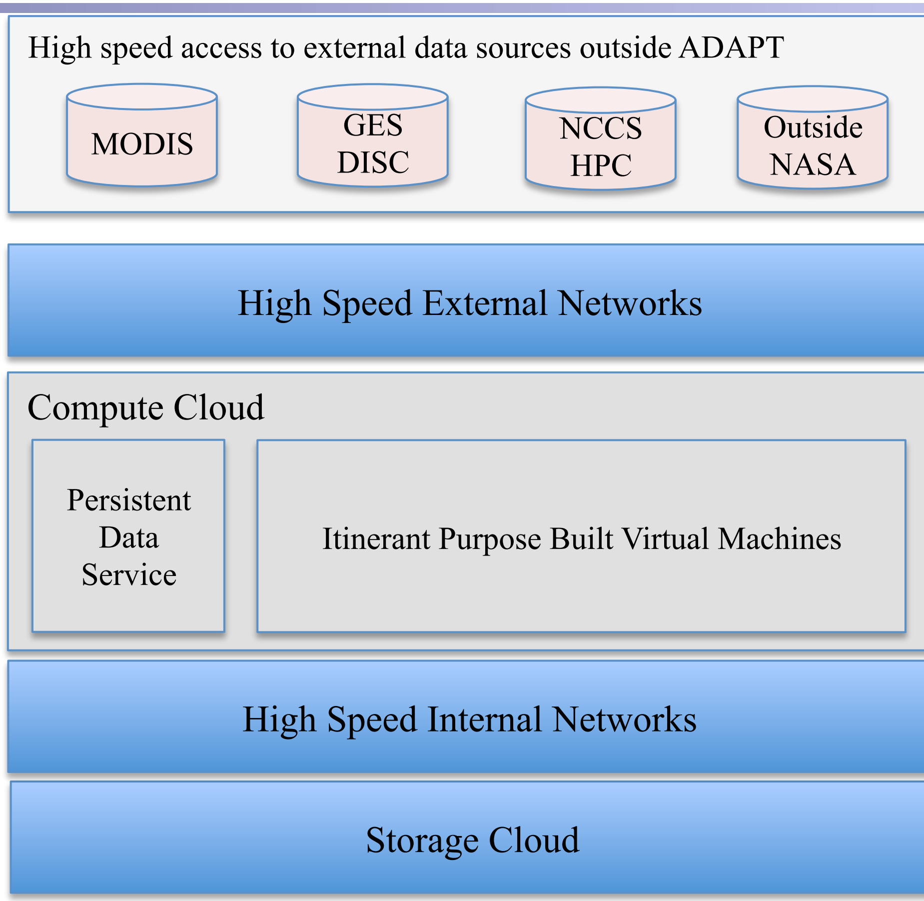
Compute systems are older, ***repurposed high performance compute nodes***

- 100s of nodes currently with plans to expand over the next 6 months
- Capable of 1,000s of virtual machines

Persistent Data Services are long lived virtual machines specifically designed for data or web services. Examples include:

- Web Portals
- Web Map Service
- FTP
- OpenDAP
- Earth System Grid Federation (ESGF)
- ESRI ArcGIS

Itinerant purpose built virtual machines are customized for each user/project. These virtual machines are not persistent and can be spun up and down as needed.



- “Data Lake” concept – storage is available as needed to all virtual environments
- Low cost, commodity based storage
- Multiple petabytes in size and easily expandable
- High performance file system using IBM GPFS

High speed external networks capable of 10 GbE and 40 GbE are available to transfer data into and out of ADAPT. In addition, remote mounts to external data sources, such as MODIS, are being served over these networks.

High speed internal networks use repurposed high performance InfiniBand switches along with more traditional Ethernet switching.

Why create the ADAPT platform?

High Performance Computing

Takes in small amounts of input and creates large amounts of output…

- Using relatively small amount of observation data, models are run to generate forecasts
- Tightly coupled processing requiring synchronization within the simulation
- Simulation applications are typically 100,000s of lines of code
- Production runs of applications push the utilization of HPC systems to be very high
- Fortran, Message Passing Interface (MPI), large shared parallel file systems
- Rigid environment – users adhere to the HPC systems

Data Analysis

Takes in large amounts of input and creates a small amount of output…

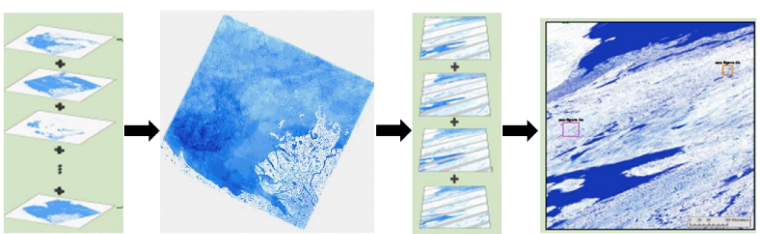
- Use large amounts of distributed observation and model data to generate science
- Loosely coupled processes requiring little to no synchronization
- Analysis applications are typically 100s of lines of code
- Require more agile development with many small runs; utilization can be low on average
- Python, IDL, Matlab, custom
- Agile environment – users run in their own environments
- Steep learning curve for these users to take advantage of HPC resources

Data Analysis is inherently different than High Performance Computing applications.

Example ADAPT Analysis Application Support

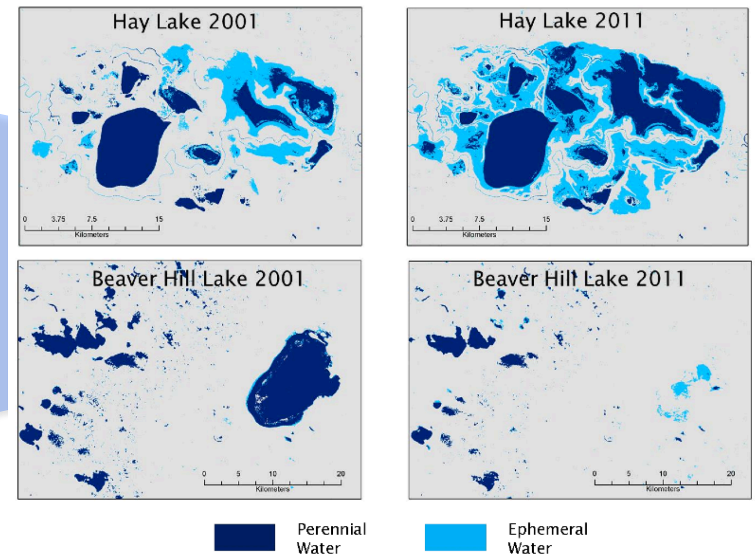
Takes in large amounts of input and creates small output

- Using large amounts of observation or model data
- Python code of 100s of lines
- Easily run in parallel across multiple virtual machines



Processing work flow for the generation of the ABoVE water maps from Landsat scenes to ABoVE tiles.

100,000 LandSat Scenes 20 TB of Data



AWM for 2001 and 2011 for Hay Lake and Beaver Hill Lake in Canada. Hay Lake has clearly expanded over this time frame while Beaver Hill Lake has diminished.

Taken from "AboVE Water Maps: 30 meter spatial resolution surface water 1991-2011," M.L. Carroll, et. al, http://above.nasa.gov/pdfs/ABoVE_water_maps_user_guide_05102016.pdf

ADAPT Status

- Operational since February of 2016
- Current users include
  - ABoVE campaign
  - CALET
  - Researchers hunting for near-Earth asteroids
  - Many others
- Access is available now for NCCS users
- Upgrading with more compute over the next few months
- Will be including user portal for self-provisioning
- Will be including capability for bursting into public clouds
  - True hybrid cloud (public/private)
- Send an email to support@nccs.nasa.gov if you would like to try out ADAPT

Data Analytics Storage Service (DASS)

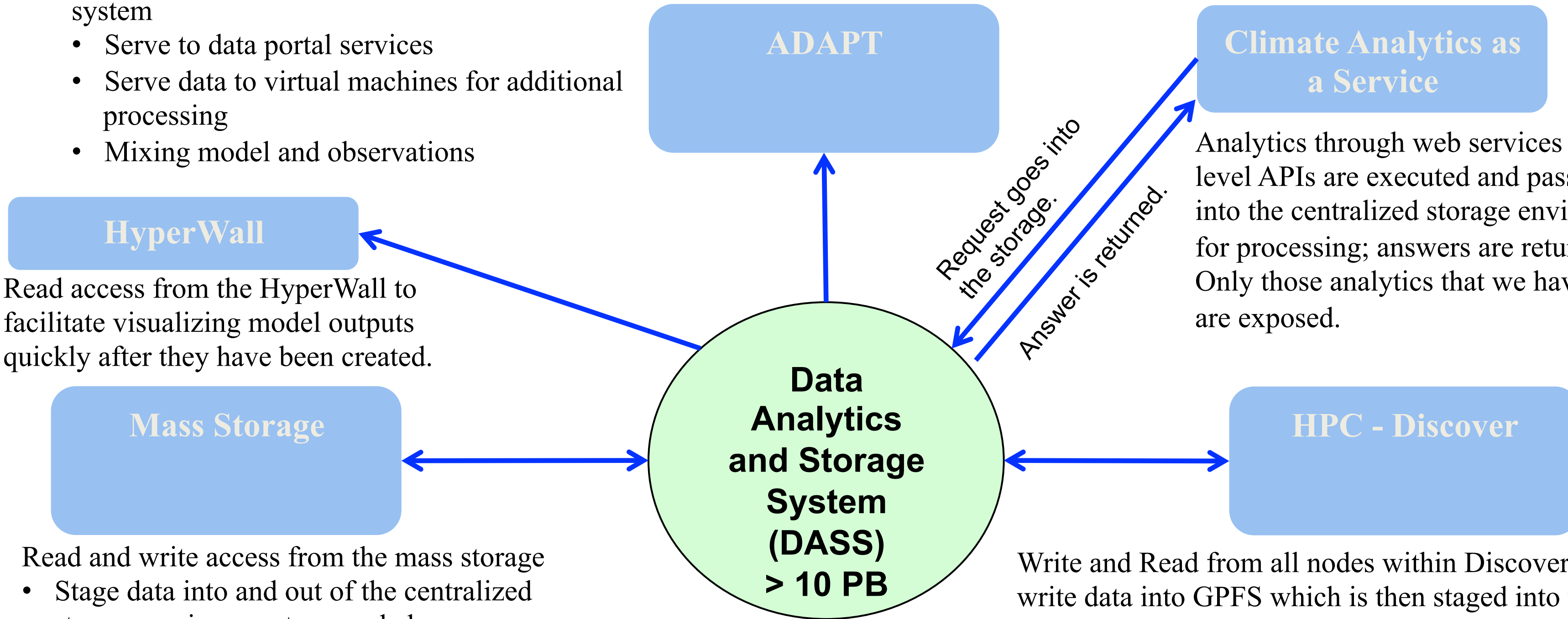
Carrie Spear,
NCCS Architect

Data Analytics Storage Service (DASS)

- Data movement and sharing of data across services within the NCCS is still a challenge
- Have large data sets created on Discover
  - On which users perform many analysis
  - And may not be in a DAAC
- Create a true centralized combination of storage and compute capability
  - Can easily share data to different services within the NCCS
  - Free up high speed disk capacity within Discover
  - Enable both traditional and emerging analytics
  - No need to modify data; use native scientific formats

DASS Concept

Read access from all nodes within the ADAPT system
- Serve to data portal services
- Serve data to virtual machines for additional processing
- Mixing model and observations



Analytics through web services or higher level APIs are executed and passed down into the centralized storage environment for processing; answers are returned. Only those analytics that we have written are exposed.

Read access from the HyperWall to facilitate visualizing model outputs quickly after they have been created.

Read and write access from the mass storage
- Stage data into and out of the centralized storage environment as needed

Write and Read from all nodes within Discover – models write data into GPFS which is then staged into the centralized storage (burst buffer like). Initial data sets could include:
- Nature Run
- Downscaling Results
- Reanalysis (MERRA, MERRA2)
- High Resolution Reanalysis

Note that more than likely all the services will still have local file systems to enable local writes within their respective security domain.

DASS Capability Overview

- 20.832 PB Raw Data Storage
- 2,604 by 8TB SAS Drives
- 14 Units
- 28 Servers
- 896 Cores
- 14,336 GB Memory
- 16 GB/Core

(3) Apollo 4520
Each Containing:
(2) ProLiant XL450
(8 each) 16GB Memory
(2 each) M.2 SSD Drives
(2 each) SSD Drives
(46) 8TB Data Drives
(6) D6000 JBODs
(70 each), 8TB Drives

(3) Apollo 4520
Each Containing:
(2) ProLiant XL450
(8 each) 16GB Memory
(2 each) M.2 SSD Drives
(2 each) SSD Drives
(46) 8TB Data Drives
(6) D6000 JBODs
(70 each), 8TB Drives

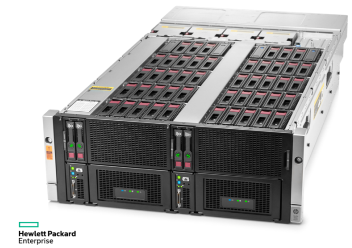
(2) HPN 5930 40GbE
32 ports each
(1) HPN 1920 1GbE
48 ports each
(2) Apollo 4520
Each Containing:
(2) ProLiant XL450
(8 each) 16GB Memory
(2 each) M.2 SSD Drives
(2 each) SSD Drives
(46) 8TB Data Drives
(4) D6000 JBODs
(70 each), 8TB Drives

(3) Apollo 4520
Each Containing:
(2) ProLiant XL450
(8 each) 16GB Memory
(2 each) M.2 SSD Drives
(2 each) SSD Drives
(46) 8TB Data Drives
(6) D6000 JBODs
(70 each), 8TB Drives

(3) Apollo 4520
Each Containing:
(2) ProLiant XL450
(8 each) 16GB Memory
(2 each) M.2 SSD Drives
(2 each) SSD Drives
(46) 8TB Data Drives
(6) D6000 JBODs
(70 each), 8TB Drives

DASS Compute/Storage Servers

- HPE Apollo 4520
  - Two (2) Proliant XL450 servers, each with
  - Two (2) 16-core Intel Haswel E5-2697Av4 2.6 GHz processors
  - 256 GB of RAM
  - Two (2) SSDs for the operating system
  - Two (2) SSDs for metadata
  - One (1) smart array P841/4G controller
  - One (1) HBA
  - One (1) Infiniband FDR/40 GbE 2-port adapter
  - Redundant power supplies
- 46 x 8 TB SAS drives
- Two (2) D6000 JBOD Shelves
- 70 x 8TB SAS drives



DASS has 14 of these units.

DASS Software Stack

Data moved from storage to compute.

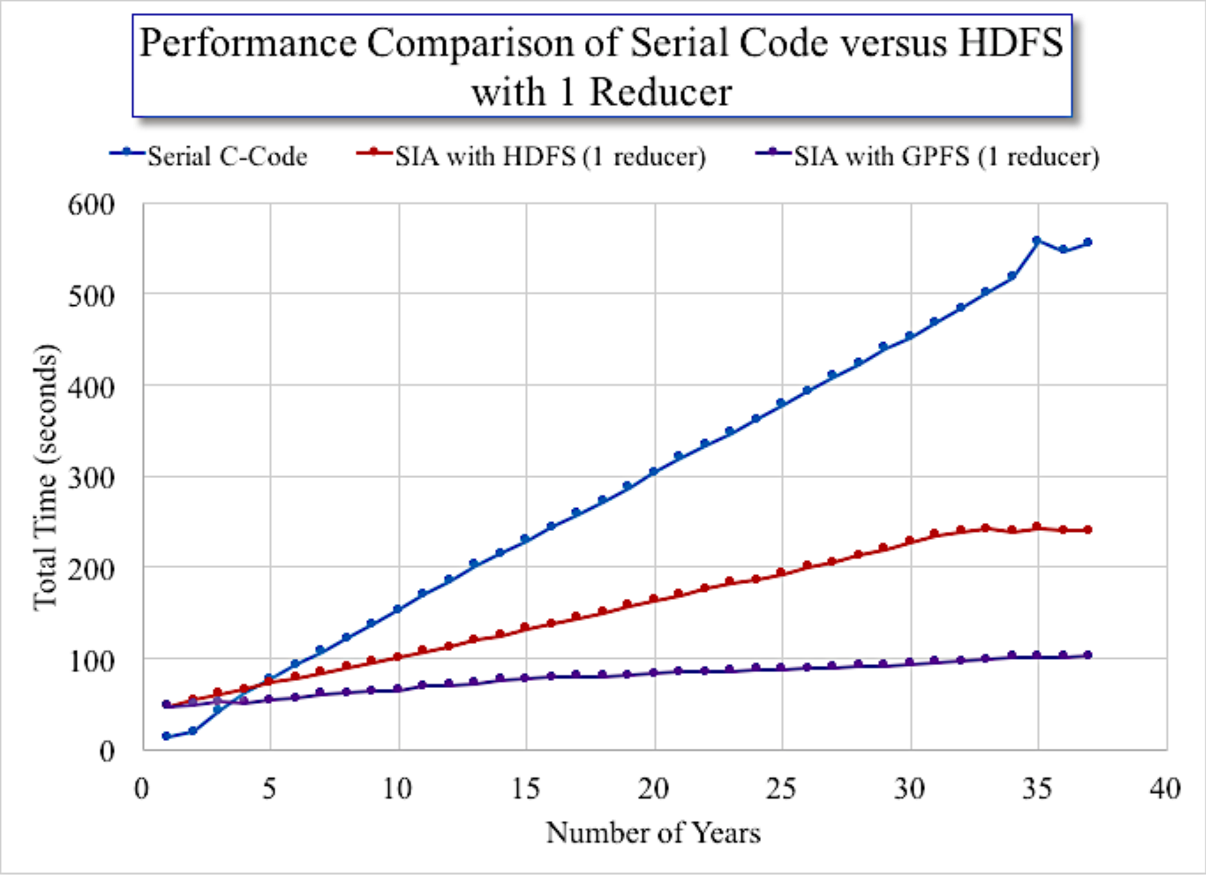
Open, Read, Write, MPI, C-code, Python, etc.	MapReduce, Spark, Machine Learning, etc.
POSIX Interface	RESTful Interface
InfiniBand, Ethernet	Cloudera and SIA
Shared Parallel File System (GPFS)	Shared Parallel File System (GPFS)

Native Scientific Data stored in
HPC Storage or
Commodity Servers and Storage

Compute moved from servers to storage.

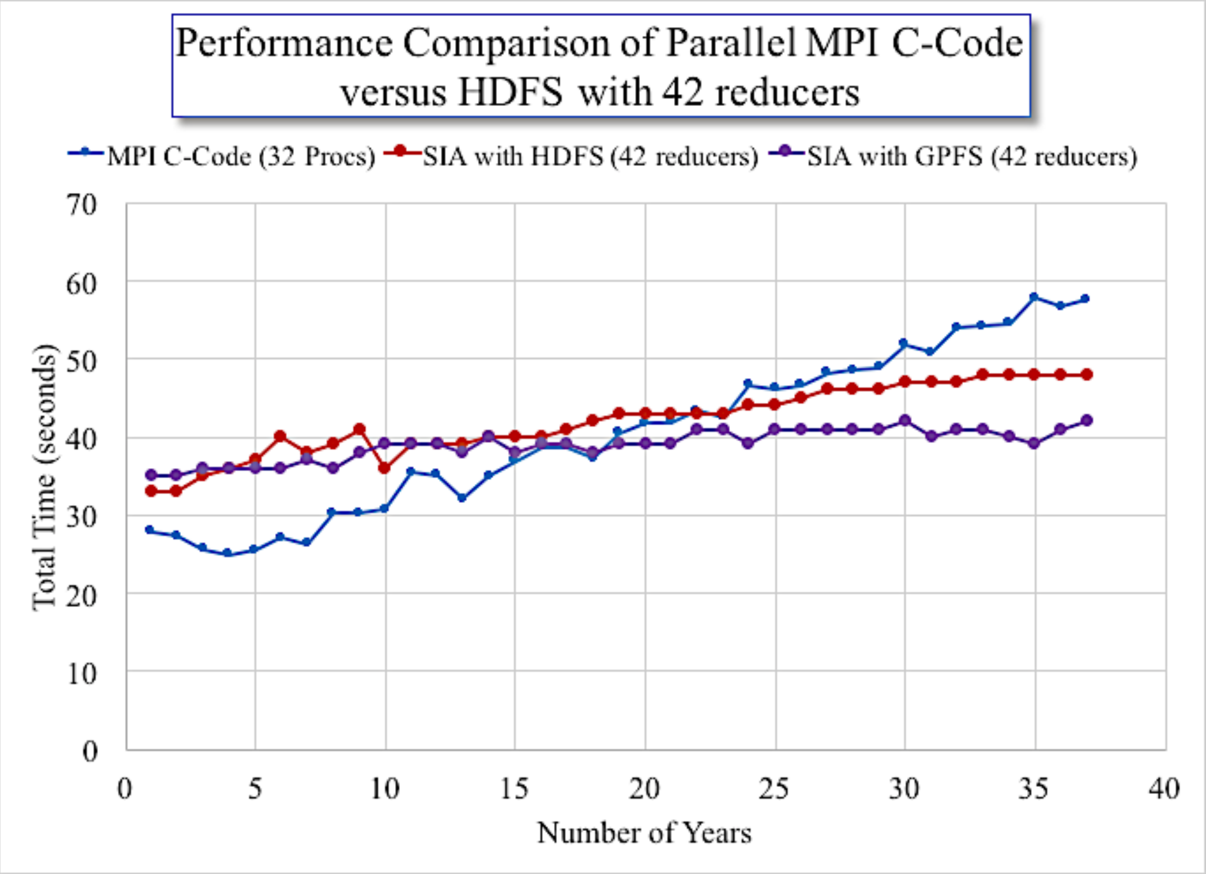
DASS: Initial Serial Performance

- Compute the average temperature for every grid point
- Vary by the total number of years
- MERRA Monthly Means
- Comparison of serial c-code to MapReduce code
- Comparison of traditional HDFS (Hadoop) where data is sequenced (modified) with GPFS where data is native NetCDF (unmodified)
- Using unmodified data in GPFS with MapReduce is the fastest

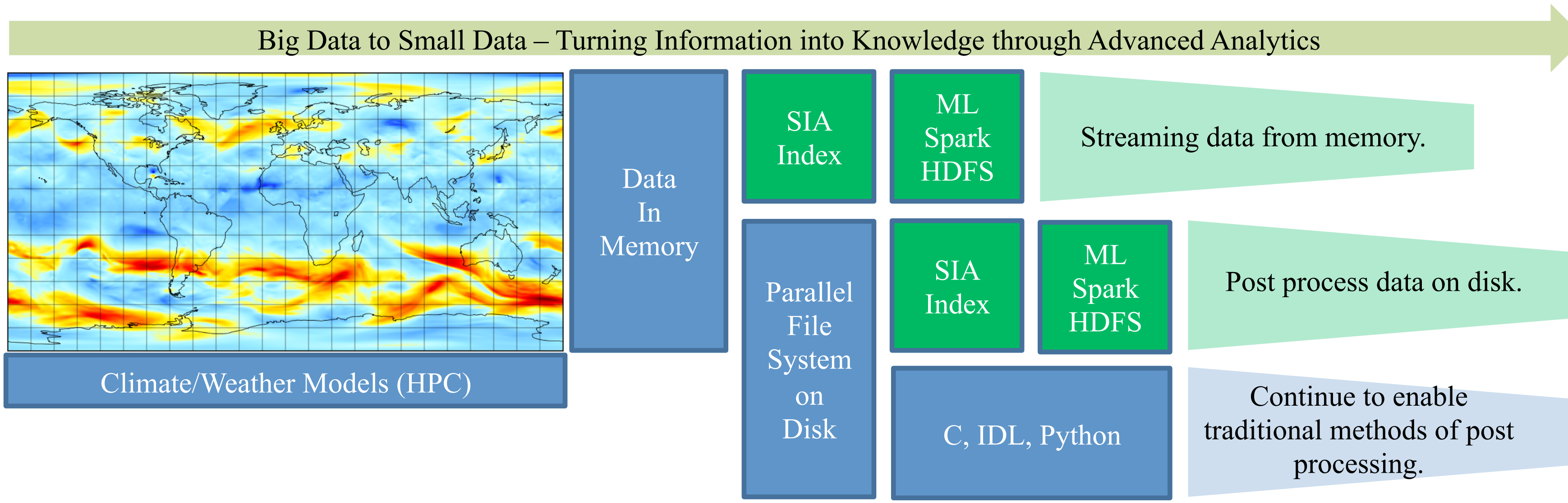


DASS: Initial Parallel Performance

- Compute the average temperature for every grid point
- Vary by the total number of years
- MERRA Monthly Means
- Comparison of serial c-code with MPI to MapReduce code
- Comparison of traditional HDFS (Hadoop) where data is sequenced (modified) with GPFS where data is native NetCDF (unmodified)
- Again using unmodified data in GPFS with MapReduce is the fastest as the number of years increases



Future of Data Analytics



- Future HPC systems must be able to efficiently transform information into knowledge using both traditional analytics and emerging machine learning techniques.
- Requires the ability to be able to index data in memory and/or on disk and enable analytics to be performed on the data where it resides
- All without having to modify the data

FY17 Hardware Updates and Procurements

Dan Duffy,
HPC Lead and NCCS Lead Architect

FY17 Plans for Discover Augmentations

- Power Expansion for Room S100
  - Additional Power Distribution Units (PDUs) and Uninterruptible Power Supply (UPS)
  - Needed capacity
  - Expect some power outages for this installation
- Discover Nobackup Disk Capacity
  - In progress
  - 13+ PB of additional capacity
  - Target operational capability January 2017

FY17 Plans for Discover Augmentations

- Expand Discover Compute Capacity
  - Additional power being installed
  - Decommission SCU9
  - Intel Skylake processors; not available until August 2017
  - Will combine FY17 and FY18 funding for a single purchase
  - Target operational capability late in 2017 early 2018
- Mass Storage Server Upgrades
  - Replacement of aging Dirac servers
  - Target operational capability by September 2017

NCCS Operational Updates

Ellen Salmon

Discover Updates

- Retiring Dali nodes, *following* upgrades to Discover login nodes
  - 256 GB memory, will be in each of ~20 Discover Haswell login nodes
  - Dali nodes will not be retired until large memory Haswell login nodes are in place
  - The 'dali<nn>' and the generic 'dali' hostnames will also be retired
  - Might need to notify your external data partners of new host names and IP addresses
- Adding a few more Gateway nodes
- Software: Rolling GPFS upgrades

JIBB (JCSDA) Migration into Discover

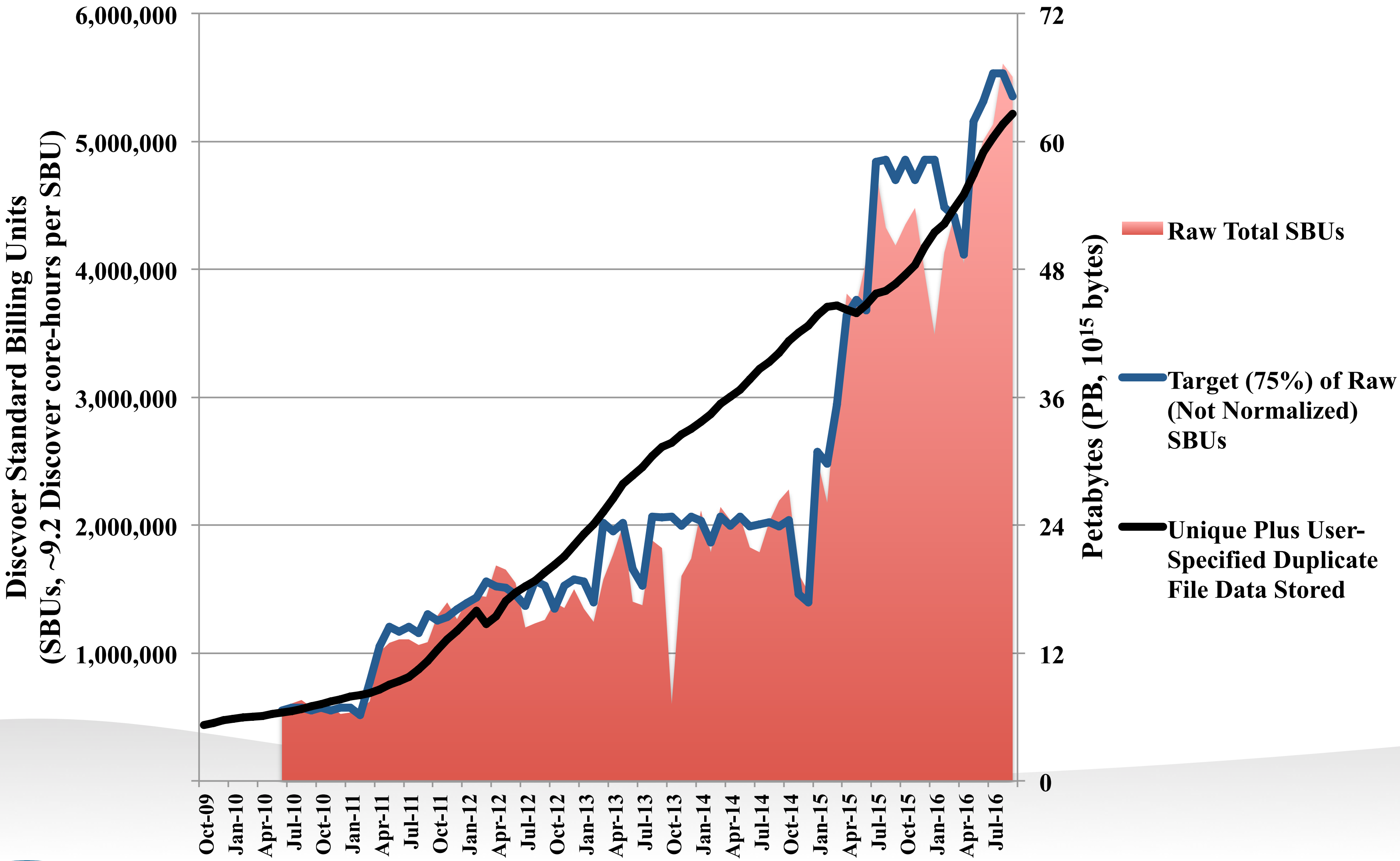
- Motivation
  - Original system installed in 2010; majority of hardware was aging, and there was no money for upgrades
  - JIBB users get access to more up-to-date processors
  - Users share lots of data between Discover and JIBB
  - Easier for the NCCS to manage one system and not two
  - Recover/re-use facilities capacity (power, floor space, and cooling)
- Status
  - Migrated active user accounts to Discover
  - Very small number of users left on JIBB, access will be cut off on the 21^{st} of this month (i.e., Friday, October 21)
  - Much of the data is copied over, but some still remains to be synchronized between Discover and JIBB

Discover NVIDIA Application Development Environment - Update

- There are currently 36 nodes in Discover with NVIDIA Graphical Processing Units (GPUs)
  - Each node contains a single NVIDIA K40 GPU
  - SandyBridge (SCU9) nodes
  - Available to everyone through the use of the following SLURM constraint:

    --constraint=k40

NCCS Total Mass Storage and Discover SBUs Targeted and Delivered



Updated NCCS Website

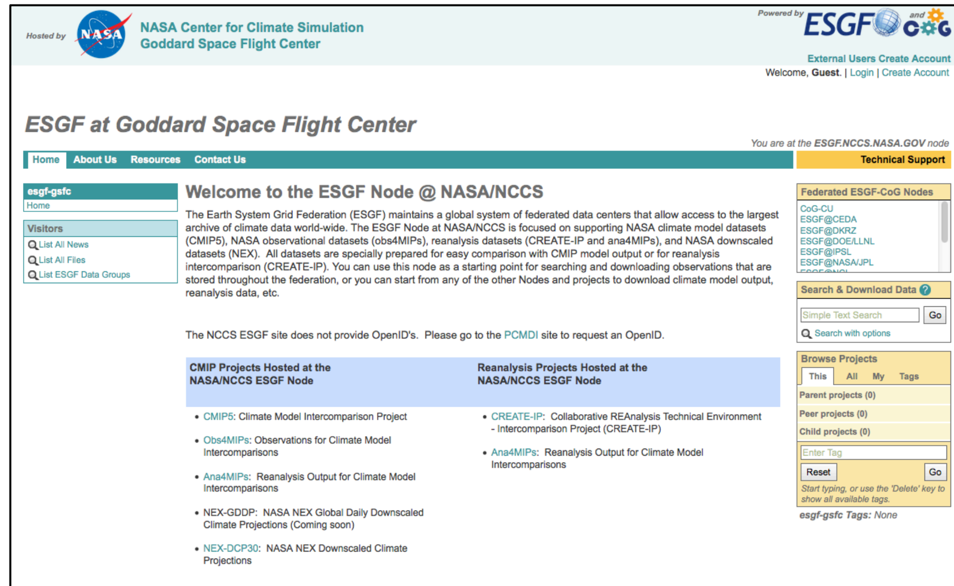
- NCCS unveiled a new look and feel for the website in late May.
  - Using Drupal for the website with plugins to allow for more rapid updates to content.
  - Much of the content has been updated, including the user information about the systems.
  - A major overhaul for the user documentation about SLURM (Discover batch scheduler).

Screen snapshot of the updated NCCS web site.

- The modernized NCCS website is easier for users to navigate, has updated information, and is a content management system that allows for more rapid additions of valuable information for the users.

ESGF Return to Service

- Earth System Grid Federation (ESGF) data portals are used to serve and download climate simulation, observation, and/or reanalysis data.
- The NASA Goddard ESGF data portal is now back online.
- Users can access and download all NASA data that was previously available.
- The NCCS team worked closely with the ESGF developers over the past year to ensure the security of the resulting ESGF node (multiple iterations of code installs, security scans, testing, etc.



Screen snapshot of the NCCS ESGF portal.

- Users of ESGF are invited to participate in a User Satisfaction Survey (https://www.surveymonkey.com/r/ESGF2016), to help ESGF developers improve the ESGF software stack and prioritize features that more closely fit the community's needs and interests.

NCCS User Recognitions

Questions & Answers

NCCS User Services:

support@nccs.nasa.gov

301-286-9120

https://www.nccs.nasa.gov

Contact Information

NCCS User Services:

support@nccs.nasa.gov

301-286-9120

https://www.nccs.nasa.gov

http://twitter.com/NASA_NCCS

Thank you

SUPPLEMENTAL SLIDES

NCCS Mass Storage October 2009 - September 2016